基于CEEMD的LSTM和ARIMA模型干旱预测适用性研究
——以新疆为例

丁　严[1]，　许德合[1]，　曹连海[1]，　管相荣[2]

（1. 华北水利水电大学测绘与地理信息学院，河南 郑州　450046；
2. 河南省自然资源电子政务中心，河南 郑州　450046）

摘　要： 干旱的频繁发生对农业生产和经济发展造成了不可忽视的危害，准确预测干旱的发生具有重要的现实意义。基于1960—2019年新疆气象站点的逐日降水量数据，计算1、3、6、9、12个月及24个月时间尺度的标准化降水指数。建立差分自回归移动平均模型（Autoregressive Integrated Moving Average，ARIMA）、长短期记忆网络（Long Short-Term Memory，LSTM）、互补集合经验模态分解（Complementary Ensemble Empirical Mode Decomposition，CEEMD）-ARIMA组合模型和CEEMD-LSTM组合模型。通过4种模型对多时间尺度SPI序列进行预测，确定各模型在干旱预测中的适用性。结果表明：（1）4种模型的预测精度均随时间尺度的增加而逐渐提高，在24个月时间尺度时达到最高；（2）CEEMD能够有效平稳时间序列，各时间尺度下，组合模型均达到了较高的预测精度，相较单一模型更适用于干旱预测；（3）4种模型预测结果精度由低到高分别为：LSTM、ARIMA、CEEMD-LSTM、CEEMD-ARIMA（决定系数最大值分别为：0.8882、0.9103、0.9403、0.9846），CEEMD-ARIMA模型相比其他3种模型效果较好，最适用于干旱预测。

关键词： 互补集合经验模态分解；长短期记忆网络；差分自回归移动平均模型；标准化降水指数；干旱预测；新疆

干旱对农业生产、经济运行、现代生活造成的危害与日俱增，也使得在气候变化过程中确保用水安全、能源安全、粮食安全变得更加困难。近百年来，中国陆地区域平均增温0.9~1.5 °C，且气温将在未来持续上升，年均降雨量虽未见显著变化，但不同区域的降雨量差异日趋明显，由此可预见大范围干旱的发生频次将会增加、强度将会增强[1-3]。随着极端天气对人类社会影响的日渐显著，如何针对极端天气的发生进行准确评估、监测和分析，成为了国内外学者关注的重点问题。

现阶段，相关研究常使用干旱指数对干旱发生的程度、持续时间和影响范围进行定量评价[4-5]。目前，学界多使用的评价指标有标准化降水指数（Standardized Precipitation Index，SPI）、帕默尔干旱指数（Plamer Drought Severity Index，PDSI）和综合干旱指数（Composite Index，CI）[6-8]。其中SPI可用于多种时间尺度下的干旱分析，干旱分级精度高且仅用降水数据即可计算，因而广泛应用于干旱研究[9-10]。降水量数据和由此计算得到的SPI具有非平稳、非线性的特征。应用这一数据进行预测，难以达到精准的预测效果。信号分解能够提取序列的局部特征并使序列平稳，国内外学者通过经验模态分解（Empirical Mode Decomposition，EMD）、集合经验模态分解（Ensemble Empirical Mode Decomposition，EEMD）、互补集合经验模态分解（Complementary Ensemble Empirical Mode Decomposition，CEEMD）对时间序列进行分解，得到了一组较为平稳的分量和一个趋势项，降低了原始时间序列的复杂度，提高

收稿日期： 2021-10-12；　**修订日期：** 2021-12-30
基金项目： 地理信息工程国家重点实验室基金（SKLGIE2019-Z-4-2）；河南省自然资源厅2020年度省自然科技项目（2020-165-10）
作者简介： 丁严（1998-），女，硕士研究生，研究方向为地理信息系统开发及应用. E-mail: 13007520896@163.com
通讯作者： 许德合. E-mail: 1445073551@qq.com

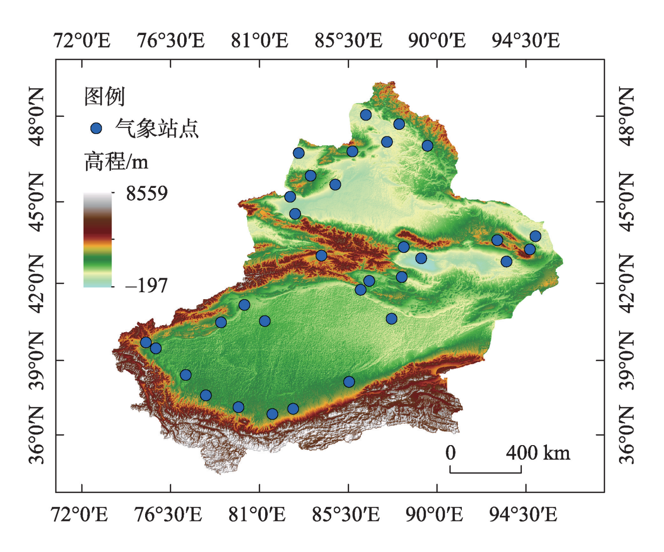
了可预测性[11-13]。在干旱预测的过程中,用于预测的模型有很多,如差分自回归移动平均模型(Autoregressive Integrated Moving Average,ARIMA)、人工神经网络(Artificial Neural Network,ANN)、支持向量机(Support Vector Machine,SVM)等,其中ARIMA模型是最常见的用于时序预测的模型[14-15]。随着机器学习的发展,长短期记忆(Long Short-Term Memory,LSTM)网络在时间序列预测中得到了应用,LSTM在处理具有很长间隔和延迟的序列上具有优势[16-17]。单一模型在时间序列的预测中容易出现局部最优的情况,预测效果不理想,因此,许多学者将信号分解与预测模型组合用于时序数据的预测,例如EMD-LSTM[18]、EEMD-ARIMA[19]、EEMD-LSTM[20]、CEEMD-LSTM[21]均得到了较好的预测结果。目前,对于组合模型预测结果适用性的评价和对比大多是组合模型与传统ARIMA模型的对比[20,22],缺乏组合模型之间的对比、组合模型与LSTM的对比。新技术新方法是否优于传统方法仍待考证。CEEMD解决了EMD模态的混叠问题以及EEMD模态的残留白噪声问题,因此,本文基于CEEMD构建CEEMD-ARIMA组合模型和CEEMD-LSTM组合模型。分别通过ARIMA、LSTM、CEEMD-ARIMA和CEEMD-LSTM模型进行预测,对其结果进行分析对比,研究其在干旱预测中的适用性。

本文选取新疆32个站点的1960—2019年逐日降水量数据,计算1、3、6、9、12个月及24个月时间尺度SPI。利用ARIMA、LSTM、CEEMD-ARIMA和CEEMD-LSTM组合模型对各SPI序列进行预测。通过对4种模型预测结果和实际计算值的对比,结合决定系数(Coefficient of Determination,R^2)、均方根误差(Root Mean Square Error,RMSE)、平均绝对误差(Mean Absolute Error,MAE)3种评价指标,分析4种模型的干旱预测精度。结合ArcGIS的经验贝叶斯克里金插值法,展示4种模型预测的干旱空间分布情况。从模型预测结果的精度和空间分布情况探索模型在干旱预测中的适用性,以期能为气象防灾减灾工作提供决策依据,减少旱灾损失。

1 数据与方法

1.1 研究区概况及数据来源

新疆地处欧亚大陆腹地,地理坐标位为73°40′~96°18′E、34°25′~48°10′N,自北向南有阿尔泰山、天山和昆仑山系,呈“三山夹两盆”的地貌格局。该区远离海洋,降水稀少,干旱频发,是典型的干旱半干旱地区。研究区域的地理位置及气象站点分布如图1所示。本文所用的逐日降水量数据来源于国家气象科学数据中心(http://data.cma.cn/)中新疆气象站观测数据。所用新疆地理高程数据来源于地理空间数据云(http://www.gscloud.cn/search)。



注:底图采用自然资源部标准地图制作,审图号为GS(2019)3333号,对底图边界无修改。下同。

图1 新疆气象站点分布

Fig. 1 Distribution of meteorological stations in Xinjiang

1.2 研究方法

1.2.1 标准化降水指数 降水量是影响干旱的重要因素。标准化降水指数考虑了降水量分布为偏态分布的情况,假定降水量分布服从Γ分布,计算出降水量的分布概率,之后进行正态标准化处理,将处理得到的结果依据气象干旱等级(GB/T20481−2017)中的干旱分级标准,进行干旱等级划分(表1)。SPI能够计算出不同时间尺度的值,满足多种水资源状况监测的需要,其中1、3、6、9、12、24个月

表1 SPI干旱分级

Tab. 1 Drought classification based on SPI

SPI范围	类型
SPI>−0.5	无旱
−1.0<SPI≤−0.5	轻旱
−1.5<SPI≤−1.0	中旱
−2.0<SPI≤−1.5	重旱
SPI≤−2.0	特旱

时间尺度下的SPI可用于描述区域的气象干旱、农业干旱、水文干旱情况[23-25]。SPI易于计算，具体计算过程参见气象干旱等级（GB/T20481-2017）。

1.2.2 CEEMD分解 1998年，Huang等[26]提出了EMD，EMD在处理非线性、非平稳信号上具有优势。原始序列输入EMD进行分解能够得到有限个固有模态函数（Intrinsic Mode Function，IMF）和趋势项，各分量包含了原始序列在不同尺度上的局部特征。经过EMD分解后的结果具有相当高的信噪比，但这种分解方法存在模态混叠的问题。EEMD作为EMD的进一步改进，通过向原始信号添加高斯白噪声，有效减少了模态混叠的发生，但白噪声的添加，使各分量含有残留白噪声[27]。Yeh等[28]提出了CEEMD，通过向原始信号中添加n组符号相反的白噪声，减少分量数据中噪声的残余量，达到残余白噪声可以忽略不计的目的，其算法步骤如下[28-29]：

（1）向原始序列$B(t)$中加入n组包括正噪声和负噪声的辅助白噪声，从而得到正噪声序列H_1和负噪声序列H_2，此时得到的序列总数为$2n$。

$$\begin{bmatrix} H_1 \\ H_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} B \\ N \end{bmatrix} \tag{1}$$

式中：N为辅助序列。

（2）将得到的序列分别进行分解，得到m个IMF分量，每组分量记为$C_{ij}^{+}(t)$和$C_{ij}^{-}(t)$，其中i=1, …, n; j=1, …, m。

（3）对每组IMF分量的$C_{ij}^{+}(t)$和$C_{ij}^{-}(t)$取平均值，得到第j个IMF的值。

$$\mathrm{IMF}_j = \frac{1}{2n}\sum_{i=1}^{n}\left[C_{ij}^{+}(t) + C_{ij}^{-}(t)\right] \tag{2}$$

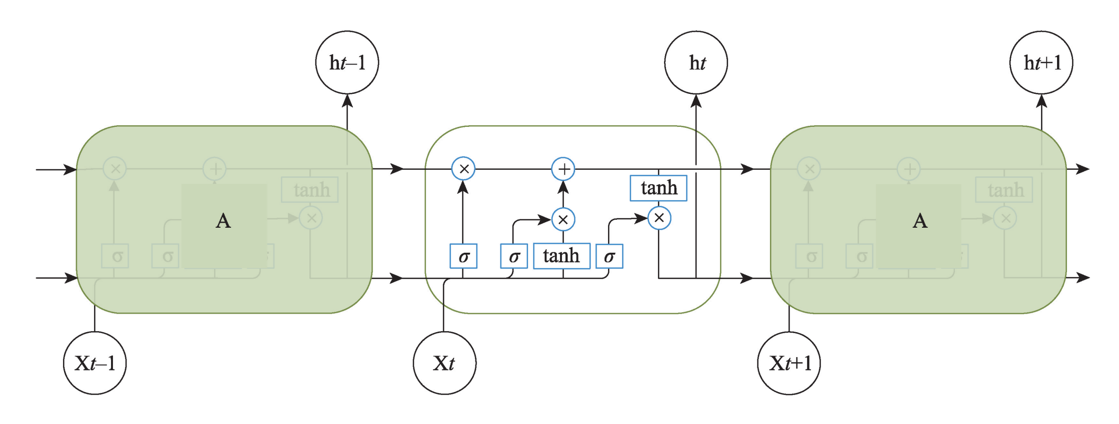
（4）将得到的IMF值作为最终分解结果，即原始序列分解为：

$$B(t) = \sum_{j=1}^{m}\mathrm{IMF}_j(t) + r(t) \tag{3}$$

式中：$r(t)$为残留趋势项。

1.2.3 LSTM网络 LSTM网络是一种特殊的循环神经网络（Recurrent Neural Network，RNN），能够学习数据传递中长期依赖的信息，并有效解决梯度问题。LSTM网络有着比RNN更复杂的重复模块（图2），其中σ、tanh分别为sigmoid函数和双曲正切函数。细胞状态是这个重复的神经网络模块链的关键，即穿过每个模块的水平线，它类似于传送带，贯穿了整个链条，保证了信息传输的不变性。通过"门"，LSTM向细胞状态添加或移除信息。遗忘门决定了要从细胞状态中移除哪些信息，这是由1个sigmoid层决定的。输入门用来更新状态信息，由两部分组成，通过sigmoid层决定哪些信息需要更新，并在tanh创建1个包含新的待添加信息的向量，由此对细胞状态进行更新。输出门用sigmoid层决定了要输出的细胞状态的部分[16]。通过运算（图2的圆圈部分），将结果继续传递给下1个单元结构。

1.2.4 ARIMA模型 Box等[30]提出了能够进行非平稳非白噪声序列预测的ARIMA模型，通过d次差分使序列平稳，然后利用自回归滑动平均（Autoregressive Moving Average，ARMA）模型预测。ARMA模型假定原始序列为一组随机序列，通过改变模型的参数对该序列近似描述，选出最符合该序列的模型参数，之后依据原始数据对未来情况进行预测[31]。ARIMA（p,d,q）模型的一般为[15]：



注：A为神经网络模块；Xt和ht分别为t时刻LSTM模块的输入和输出。

图2 LSTM结构图

Fig. 2 Structure diagram of LSTM

$$Y_t = \omega_1 Y_{t-1} + \omega_2 Y_{t-2} + \cdots + \omega_p Y_{t-p} + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \cdots - \theta_q u_{t-q} \quad (4)$$

式中：Y_t 为时间序列值；$\omega_i (i=1,2,\cdots,p)$ 和 $\theta_j (j=1,2,\cdots,q)$ 分别为自回归系数和滑动平均系数；u_t 为白噪声序列，且 $u_t \sim N(0,\sigma^2)$。

ARIMA模型的建模流程为：

（1）平稳性检验。本文通过单位根检验（Augmented Dickey-Fuller Test，ADF）判断时间序列的平稳性[32]。若为非平稳时间序列则需对原始序列 d 次差分。

（2）确定模型阶数的取值范围。根据数据的自相关函数（Autocorrelation Function，ACF）和偏自相关函数（Partial Autocorrelation Function，PACF）确定 p、q 的取值范围。

（3）模型定阶。利用赤池信息准则（Akaike Information Criterion，AIC）、贝叶斯信息准则（Bayesian Information Criterion，BIC）对模型定阶，AIC、BIC公式如下：

$$\mathrm{AIC}(p,q) = N\ln\sigma^2(p,q) + 2(p+q+1)$$
$$\mathrm{BIC}(p,q) = N\ln\sigma^2(p,q) + (p+q+1)\ln N \quad (5)$$

式中：N 为参数个数。选择AIC、BIC值最小时对应的 p、q 值。

1.2.5　基于CEEMD的组合模型　波动性强的原始序列经过CEEMD分解，能够得到一组波动较低的IMF分量，这提高了序列的可预测性。通过Python，将CEEMD分别与LSTM和ARIMA模型结合组成CEEMD-LSTM组合模型和CEEMD-ARIMA组合模型。通过组合模型进行预测的步骤如下：

（1）CEEMD分解。通过CEEMD对原始SPI序列进行分解，得到从高频到低频的IMF1、IMF2、···、IMFn 以及Res。

（2）LSTM或ARIMA模型预测。将IMF1、IMF2、···、IMFn 以及Res分别导入LSTM或ARIMA模型进行预测，预测结果分别记为 $P1$、$P2$、···、$Pn+1$。

（3）对预测结果相加求和。

$$P = \sum_{i=1}^{n+1} P_i \quad (6)$$

基于CEEMD的组合模型建模流程如图3所示。

1.2.6　评价指标　本文选取RMSE、MAE、R^2 作为4种模型的评价指标。RMSE和MAE的取值范围为[0，+∞]，值越小，模型效果越好。R^2 越大，表示拟合效果越好，最大值为1。

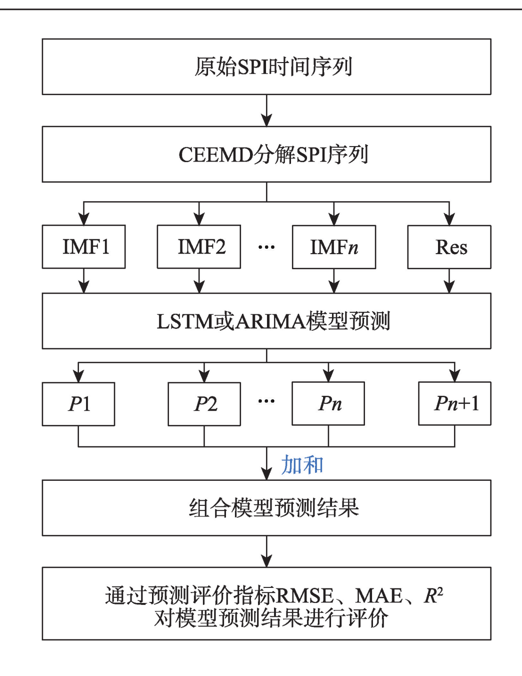


图3　组合模型建立流程

Fig. 3　Workflow of combined model

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2} \quad (7)$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \hat{y}_i\right| \quad (8)$$

$$R^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2 - \sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (9)$$

式中：x_i 是观测值；y_i 是真实值；$\bar{y}$ 是 y_i 的平均值；$\hat{y}_i$ 为预测值；N 为样本数。

2　结果与分析

2.1　LSTM网络模型训练及预测

本文以库尔勒站点为例，利用LSTM网络模型对1、3、6、9、12个月及24个月时间尺度SPI序列进行建模，步骤如下：

（1）数据归一化处理

对输入的SPI数据进行归一化处理，以提高模型的训练速度。

（2）网络模型训练

LSTM网络的激活函数通常有sigmoid、tanh和ReLU。sigmoid存在着随神经网络层数加深，梯度后向传播到浅层网络时易出现梯度消失的缺点；tanh也存在梯度消失的情况，且sigmoid和tanh的随机梯度下降收敛速度较慢，因此激活函数选用了

ReLU。1次训练选取的样本数为1，即每训练1个样本，更新1次权重。损失函数则采用均方误差（Mean Squared Error，MSE），优化算法采用了Adam。通过"早停法"防止训练过拟合，即随着迭代次数增加，MSE逐渐下降，模型精度逐渐提高；当MSE值上升时，停止训练。为确保模型精度达到最高，迭代次数设置为300。采用了黄金分割法选择隐藏神经元数量，隐藏层神经元数为25[17]。

（3）输出预测数据

由于之前对数据进行了归一化处理，因此，此处需要采取反归一化处理，以得到模型的实际预测数据（图4）。

2.2 ARIMA模型建模及预测

依据32个气象站点1960—2019年的逐日降水量数据进行SPI计算。不同时间尺度的SPI适用于干旱研究的不同方面，因此本文计算了1、3、6、9、12、24个月共6个时间尺度的SPI。将计算得到的SPI中1960—2007年数据作为训练集，2008—2019年数据作为测试集。本文以库尔勒站点为例对ARIMA建模，在预测前，需要对测试集数据的平稳性进行判断。若数据平稳，则可通过ARMA模型进行预测；若不平稳，则需进行差分，ADF检验结果见表2。表2中6个时间尺度SPI的P值均小于0.05，即时间序列均为平稳时间序列，因此，可进行下一步。

通过ACF、PACF确定各时间序列p、q的可能取值。利用AIC、BIC准则选取最优模型。各序列的模型定阶结果见表3。分别通过6个时间尺度SPI的最优模型进行预测，预测结果见图4。

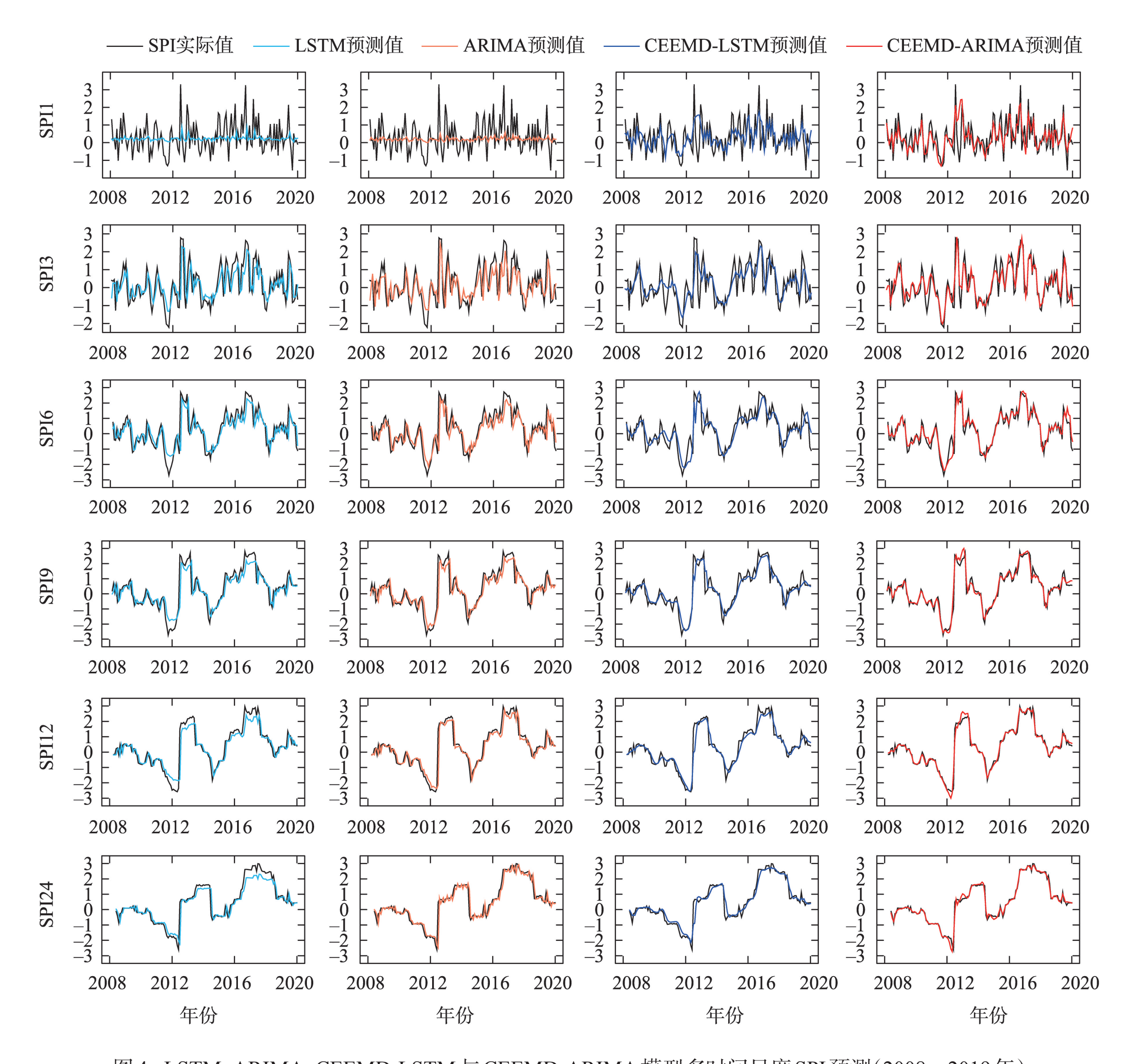


图4　LSTM、ARIMA、CEEMD-LSTM与CEEMD-ARIMA模型多时间尺度SPI预测（2008—2019年）

Fig. 4　Forecast of multi-time scale SPI of LSTM, ARIMA, CEEMD-LSTM and CEEMD-ARIMA model (2008–2019)

表2 原始序列单位根检验

Tab. 2 ADF test of the original sequence

SPI序列	单位根检验	临界值			*P*值
		1%	5%	10%	
SPI1	−8.0407	−3.4419	−2.8667	−2.5695	1.8500e−12
SPI3	−9.4801	−3.4419	−2.8666	−2.5695	3.8938e−16
SPI6	−6.7711	−3.4420	−2.8667	−2.5695	2.6407e−09
SPI9	−4.4529	−3.4423	−2.8668	−2.5696	0.0002
SPI12	−4.0259	−3.4423	−2.8668	−2.5696	0.0013
SPI24	−3.8011	−3.4425	−2.8669	−2.5696	0.0029

表3 6个尺度SPI的ARIMA模型定阶

Tab. 3 ARIMA model order based on SPI values of six time scales

SPI序列	*p*	*d*	*q*	AIC	BIC
SPI1	1	0	0	1752.561	1766.295
SPI3	0	0	2	1511.110	1529.410
SPI6	4	0	1	1274.699	1306.696
SPI9	2	0	1	1044.266	1067.099
SPI12	2	0	1	648.904	671.716
SPI24	9	0	2	181.161	240.251

2.3 利用组合模型对SPI序列进行预测

经过参数的多次修改和对比，最终选定将Nstd设置为0.2，NE设置为100，TNM设置为8。利用CEEMD分解多尺度SPI，得到8个IMF分量和1个趋势项。以SPI3分解为例，原始序列和分解得到的子序列见图5。由图5可知，原始序列波动范围较大，而分解得到的IMF分量波动范围较小，随着分解的逐步进行，分量的波动趋于平缓，说明通过CEEMD分解能够降低原始序列的非平稳性。

选取1960—2007年数据作为训练集，2008—2019年数据作为测试集。利用组合模型进行预测，预测结果见图4。由图4可知，在1个月时间尺度下，LSTM和ARIMA模型的预测值与实际观测计算值相差较大。CEEMD-LSTM和CEEMD-ARIMA组合模型的预测值与实际值则较接近，其中CEEMD-ARIMA能准确预测到2011年的干旱发生强度。在3个月尺度下，2个单一模型的预测值与实际值差距缩小，预测的SPI变化趋势与实际趋势相符。此时，CEEMD-ARIMA已能准确预测2011年和2017年的干旱情况，整体预测结果与实际情况最为一致。在6个月尺度下，与LSTM相比，ARIMA模型对干旱发生时间和强度的预测更为准确。4个模型中，ARIMA和CEEMD-ARIMA模型对干旱的预测较精准。在9个月尺度和12个月尺度下，除LSTM外的其他3种模型预测情况接近实际情况，较1、3、6个月尺度下，对干旱事件的发生及强度和持续时间的预测更为准确。在24个月尺度下，4种模型的预测结果与

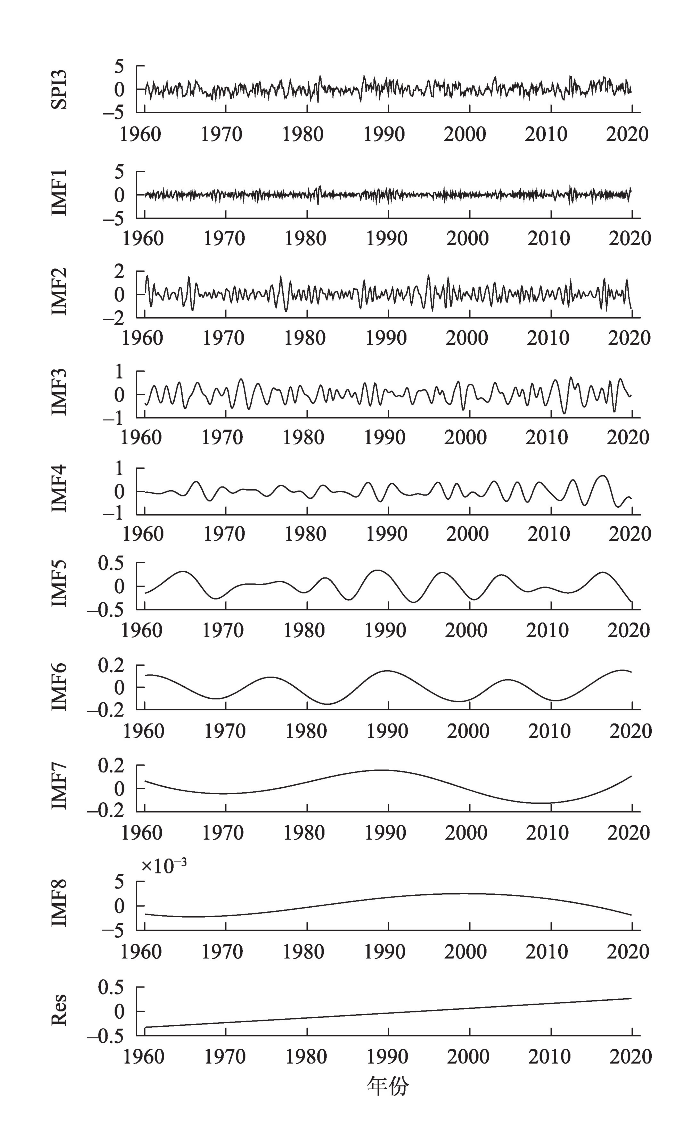


图5 CEEMD分解SPI3序列

Fig. 5 CEEMD decomposition results of SPI3 sequence

实际情况近乎一致，从干旱发生强度的预测情况来看，ARIMA和CEEMD-ARIMA模型的预测结果分别优于LSTM和CEEMD-LSTM模型。对模型在6个时间尺度SPI的预测结果进行对比，在1个月时间尺度下，4种模型的预测结果均为6个时间尺度中最差的，与实际结果相差最大。随着时间尺度的增大，4种模型预测的准确性有所提升。

通过R^2、RMSE、MAE共3种评价指标对预测结果进行评价，进一步分析4种模型的预测精度。表4中LSTM在SPI1的RMSE、MAE值分别为0.8681和0.6478，随着时间尺度的增加RMSE、MAE值逐渐减小。在24个月时间尺度下达到最小，SPI24的RMSE、MAE值分别为0.4266和0.2700。R^2值则呈现相反趋势，表明随着时间尺度增大，模型的预测精度逐渐提高。ARIMA、CEEMD-ARIMA、CEEMD-LSTM模型预测精度随时间尺度的变化趋势与LSTM一致。对各时间尺度SPI进行预测，ARIMA模型预测结果的R^2值均略高于LSTM，RMSE、MAE的值则均略低于LSTM，说明ARIMA模型的预测精度优于LSTM。CEEMD-LSTM和CEEMD-ARIMA模型的R^2值在各时间尺度均高于单一模型，LSTM、ARIMA、CEEMD-LSTM和CEEMD-ARIMA模型在SPI24的R^2值分别为0.8882、0.9103、0.9403和0.9846。其中，CEEMD-ARIMA模型除对SPI1的预测结果外，R^2值均在0.8以上，具有较高的预测精度。在各个时间尺度下，预测精度从低到高为：LSTM、ARIMA、CEEMD-LSTM、CEEMD-ARIMA模型，说明ARIMA的预测精度高于LSTM，CEEMD能够有效提高模型的预测精度。

表4　4种模型预测结果的R^2、RMSE、MAE值

Tab. 4　R^2, RMSE and MAE values of the predicted results of four models

时间尺度	模型	R^2	RMSE	MAE
1个月	LSTM	−0.0146	0.8681	0.6478
	ARIMA	−0.0058	0.8643	0.6431
	CEEMD-LSTM	0.2648	0.7389	0.5683
	CEEMD-ARIMA	0.4488	0.6398	0.4828
3个月	LSTM	0.4200	0.7906	0.6040
	ARIMA	0.4986	0.7350	0.5531
	CEEMD-LSTM	0.5782	0.6742	0.5017
	CEEMD-ARIMA	0.8246	0.4347	0.3355
6个月	LSTM	0.6686	0.6595	0.4710
	ARIMA	0.6870	0.6410	0.4554
	CEEMD-LSTM	0.7776	0.5402	0.4116
	CEEMD-ARIMA	0.9153	0.3334	0.2397
9个月	LSTM	0.7873	0.5732	0.3856
	ARIMA	0.8039	0.5503	0.3553
	CEEMD-LSTM	0.8921	0.4082	0.2839
	CEEMD-ARIMA	0.9619	0.2426	0.1789
12个月	LSTM	0.8592	0.4858	0.3084
	ARIMA	0.8732	0.4610	0.2628
	CEEMD-LSTM	0.9302	0.3420	0.2251
	CEEMD-ARIMA	0.9793	0.1863	0.1271
24个月	LSTM	0.8882	0.4266	0.2700
	ARIMA	0.9103	0.3822	0.2109
	CEEMD-LSTM	0.9403	0.3119	0.1958
	CEEMD-ARIMA	0.9846	0.1584	0.1019

使用ArcGIS对32个站点在2019年SPI的实际观测计算值和预测值进行可视化展示(图6)。由于新疆的干旱在一年四季皆有发生，此处选择能够进行降雨量季节变化分析的SPI3对区域干旱情况进行展示。从图6中可以看出，CEEMD-ARIMA组合模型对干旱空间分布情况的预测与实际情况最为接近。2019年2月的北疆降水量偏多，全疆其余大部分偏少。4种模型在冬季的预测情况与实际情况都存在着偏差，其中CEEMD-ARIMA组合模型的预测结果与实际计算结果较为一致。

3　讨　论

SPI时间序列是非平稳序列，而单一模型预测结果的精度受原始数据平稳性影响较大。Liu等[33]利用ARMA对山东省5个站点的SPI9序列进行预测，预测结果的平均相对误差最低为20.39%，最高为43.69%，预测精度较低且不同站点间存在很大差异。单独通过LSTM预测SPI，同样有着较差的预测结果[34]。CEEMD分解能够为模型预测提供平稳性，从而提高序列的可预测性[13]。通过CEEMD分解，原始序列在不同尺度的局部特征被提取出来，非平稳时间序列转化为平稳的分量。因此，本研究利用CEEMD降低SPI序列的非平稳性，确保LSTM和ARIMA模型能够有效预测SPI序列。

在4种模型的预测结果中，SPI1的预测精度相较于其他5个时间尺度最差。数据的平稳性与预测结果有密切关系，1个月时间尺度的数据量是6个时间尺度中最大的，并且数据序列趋于严平稳(序列的分布结构不随时间改变)，随着时间尺度的增大，数据量减少，并且数据序列趋于宽平稳(未来值与

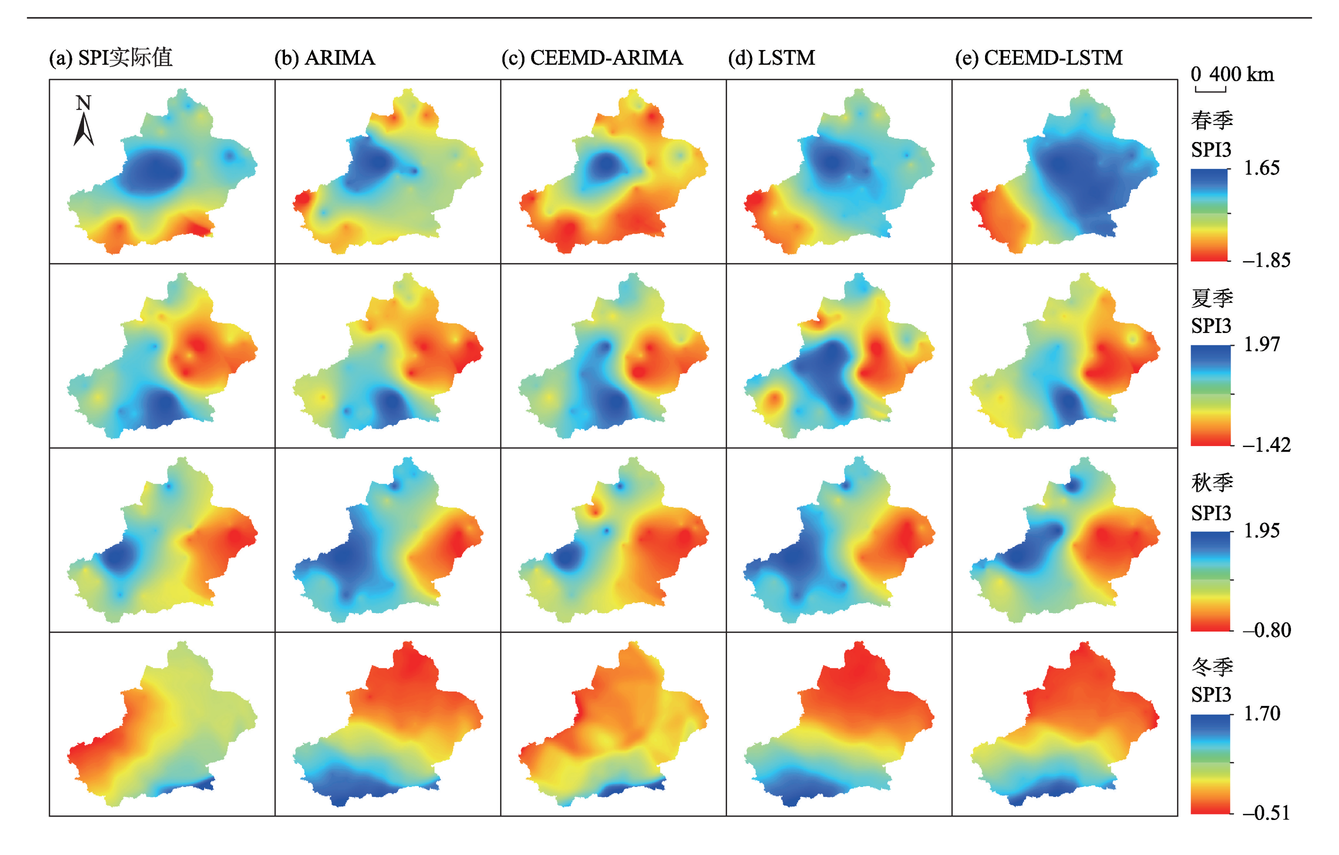


图6 使用克里金插值对实际值和4种模型的预测结果可视化展示

Fig. 6 Kriging interpolation results of the actual calculated values and the predicted values of four models

过去值相关),模型的预测情况变好。LSTM在高频序列的预测中具有较高的预测精度,ARIMA在低频序列中有较好的预测效果,因此,LSTM和ARIMA模型分别适用于高频序列和低频序列的预测,同时也造成了LSTM在SPI序列预测中预测效果略差于ARIMA模型[16]。CEEMD分解得到的子序列可预测性要高于原始序列,因此,在1个月时间尺度下,2个组合模型的预测情况明显优于单一模型的预测情况。在3个月和6个月时间尺度下,组合模型优于单一模型。随着时间尺度的增大,优势逐渐缩小,长时间尺度的SPI序列集合了原始数据中更多的信息,整个序列趋于平稳,单一模型的预测精度随之提高。

SPI易于计算,且能够描述地区的气象干旱、农业干旱、水文干旱情况,但对于新疆这一研究区而言,SPI具有一定的局限性。新疆农业所耗水分不仅来源于降水,也来源于当地的灌溉用水。地下水位的变化与山区河流径流及新疆农业耗水有着很大的关系。干旱的发生是多种因素的共同作用,除降水外,需要考虑的因素还有很多。在年降水量未有显著变化的情况下,随着全球温度的逐渐上升,干旱发生的频次势必会增加。因此,若只考虑降水因素的影响,干旱发生的预测将会变得越来越困难,还需在研究中考虑多种因素的干旱指数在干旱预测中的适用性。

4 结 论

本文分别利用LSTM、ARIMA、CEEMD-LSTM和CEEMD-ARIMA模型对1、3、6、9、12个月及24个月时间尺度的SPI进行预测,通过对预测结果的对比分析,主要得到以下结论:

(1) 4种模型预测精度随时间尺度的增大而提高,即在1个月尺度下最低,在24个月尺度下最高,此时R^2值均在0.85以上,表明4种模型在干旱预测中的适用性随着时间尺度的增大逐渐提高。

(2) CEEMD-LSTM和CEEMD-ARIMA组合模型在1、3、6、9、12个月及24个月时间尺度下,均有着比单一模型更高的精度。说明CEEMD在处理非平稳、非线性数据上具有优势,通过CEEMD分解,原始数据序列变得平稳,序列的可预测性提高。

(3) CEEMD-ARIMA模型的预测精度最高,除SPI1外,其余5个时间尺度的R^2值均在0.80以上,且在SPI24时达到了0.98。CEEMD-ARIMA模型预测

的干旱空间分布情况与实际情况较为吻合,说明CEEMD-ARIMA模型能够很好地拟合不同尺度的SPI序列,适用于干旱预测。

参考文献(References):

[1] 李夫鹏, 王正涛, 超能芳, 等. 利用Swarm星群探测亚马逊流域2015—2016年干旱事件[J]. 武汉大学学报·信息科学版, 2020, 45(4): 595-603. [Li Fupeng, Wang Zhengtao, Chao Nengfang, et al. 2015-2016 drought event in the Amazon River basin as measured by Swarm constellation[J]. Geomatics and Information Science of Wuhan University, 2020, 45(4): 595-603.]

[2] 田丰, 武建军, 刘雷震, 等. 1901—2015年华北平原干旱时空转移特征及热点区域探测[J]. 干旱区资源与环境, 2020, 34(6): 87-96. [Tian Feng, Wu Jianjun, Liu Leizhen, et al. Spatiotemporal transferring characteristics of drought and its hotpots detection in North China Plain during 1901-2015[J]. Journal of Arid Land Resources and Environment, 2020, 34(6): 87-96.]

[3] 卢宝宝, 孙慧兰, 姜泉泉, 等. 近53 a新疆水分盈亏量时空变化特征[J]. 干旱区研究, 2021, 38(6): 1579-1589. [Lu Baobao, Sun Huilan, Jiang Quanquan, et al. Spatiotemporal variation characteristics of the water budget in Xinjiang during the latest 53 years[J]. Arid Zone Research, 2021, 38(6): 1579-1589.]

[4] 宋玉鑫, 左其亭, 马军霞. 基于SWAT模型的开都河流域水文干旱变化特征及驱动因子分析[J]. 干旱区研究, 2021, 38(3): 610-617. [Song Yuxin, Zuo Qiting, Ma Junxia. Variation and dynamic drivers of drought in Kaidu River Basin based on the SWAT model[J]. Arid Zone Research, 2021, 38(3): 610-617.]

[5] Vasiliades L, Loukas A, Liberis N. A water balance derived drought index for Pinios River Basin, Greece[J]. Water Resources Management, 2011, 25(4): 1087-1101.

[6] 方秀琴, 郭晓萌, 袁玲, 等. 随机森林算法在全球干旱评估中的应用[J]. 地球信息科学学报, 2021, 23(6): 1040-1049. [Fang Xiuqin, Guo Xiaomeng, Yuan Ling, et al. Application of random forest algorithm in global drought assessment[J]. Journal of Geo-Information Science, 2021, 23(6): 1040-1049.]

[7] 刘媛媛, 李霞, 王小博, 等. 2001—2018年中国-老挝交通走廊核心区植被稳定性对极端干旱的响应[J].生态学报, 2021, 41(7): 2537-2547. [Liu Yuanyuan, Li Xia, Wang Xiaobo, et al. Vegetation stability in response to extreme droughts from 2001 to 2018 in the core area of China-Laos transportation corridors[J]. Acta Ecologica Sinica, 2021, 41(7): 2537-2547.]

[8] 周丽, 谢舒蕾, 吴彬. 基于CI和强度分析方法的四川冬春季干旱事件变化特征[J]. 自然灾害学报, 2020, 29(3): 36-44. [Zhou Li, Xie Shulei, Wu Bin. Variation characteristics of the winter and spring drought events in Sichuan based on CI and Intensity analysis[J]. Journal of Natural Disasters, 2020, 29(3): 36-44.]

[9] 徐一丹, 任传友, 马熙达, 等. 基于SPI/SPEI指数的东北地区多时间尺度干旱变化特征对比分析[J].干旱区研究, 2017, 34(6): 1250-1262. [Xu Yidan, Ren Chuanyou, Ma Xida, et al. Change of drought at multiple temporal scales based on SPI /SPEI in Northeast China[J]. Arid Zone Research, 2017, 34(6): 1250-1262.]

[10] 李明, 葛晨昊, 邓宇莹, 等. 黄土高原气象干旱和农业干旱特征及其相互关系研究[J]. 地理科学, 2020, 40(12): 2105-2114. [Li Ming, Ge Chenhao, Deng Yuying, et al. Meteorological and agricultural drought characteristics and their relationship across the Loess Plateau[J]. Scientia Geographica Sinica, 2020, 40(12): 2105-2114.]

[11] Rezaei H, Faaljou H, Mansourfar G. Stock price prediction using deep learning and frequency decomposition[J]. Expert Systems with Applications, 2020, 169(12): 114332, doi: 10.1016/j.eswa.2020.114332.

[12] 徐岩岩, 常军. 基于DERF2.0模式1～52天最低温度逐日预报的检验评估[J] .高原气象, 2018, 37(4): 1042-1050. [Xu Yanyan, Chang Jun. Evaluation of the minimum temperature forecast of 1~52 days based on DERF2.0 model[J]. Plateau Meteorology, 2018, 37(4): 1042-1050.]

[13] Wu C, Wang J, Chen X, et al. A novel hybrid system based on multi-objective optimization for wind speed forecasting[J]. Renewable Energy, 2020, 146(8): 149-165.

[14] 王蕾, 王鹏新, 李俐, 等. 应用条件植被温度指数预测县域尺度小麦单产[J]. 武汉大学学报·信息科学版, 2018, 43(10): 1566-1573. [Wang Lei, Wang Pengxin, Li Li, et al. Wheat yield forecasting at county scale based on time series vegetation temperature condition index[J]. Geomatics and Information Science of Wuhan University, 2018, 43(10): 1566-1573.]

[15] 杨慧荣, 张玉虎, 崔恒建, 等. ARIMA和ANN模型的干旱预测适用性研究[J]. 干旱区地理, 2018, 41(5): 945-953. [Yang Huirong, Zhang Yuhu, Cui Hengjian, et al. Applicability of ARIMA and ANN models for drought forecasting[J]. Arid Land Geography, 2018, 41(5): 945-953.]

[16] Liu M D, Ding L, Bai Y L. Application of hybrid model based on empirical mode decomposition, novel recurrent neural networks and the ARIMA to wind speed prediction[J]. Energy Conversion and Management, 2021, 233: 113917, doi: 10.1016/j.enconman.2021.113917.

[17] 张建海, 张棋, 许德合, 等. ARIMA-LSTM组合模型在基于SPI干旱预测中的应用——以青海省为例[J].干旱区地理, 2020, 43(4): 1004-1013. [Zhang Jianhai, Zhang Qi, Xu Dehe, et al. Application of a combined ARIMA-LSTM model based on SPI for the forecast of drought: A case study in Qinghai Province[J]. Arid Land Geography, 2020, 43(4): 1004-1013.]

[18] 王亦斌, 孙涛, 梁雪春, 等. 基于EMD-LSTM模型的河流水量水位预测[J]. 水利水电科技进展, 2020, 40(6): 40-47. [Wang Yibin, Sun Tao, Liang Xuechun, et al. Prediction of river water flow and water level based on EMD-LSTM model[J]. Advances in Science and Technology of Water Resources, 2020, 40(6): 40-47.]

[19] 刘艳, 杨耘, 聂磊, 等. 玛纳斯河出山口径流EEMD-ARIMA预测[J]. 水土保持研究, 2017, 24(6): 273-280, 285. [Liu Yan, Yang

Yun, Nie Lei, et al. The EEMD-ARIMA prediction of runoff at mountain pass of Manas River[J]. Research of Soil and Water Conservation, 2017, 24(6): 273–280, 285.]

[20] 杨倩, 秦莉, 高培, 等. 基于EEMD-LSTM模型的天山北坡经济带年降水量预测[J]. 干旱区研究, 2021, 38(5): 1235-1243. [Yang Qian, Qin Li, Gao Pei, et al. Prediction of annual precipitation in the northern slope economic belt of Tianshan Mountains based on a EEMD-LSTM model[J]. Arid Zone Research, 2021, 38 (5): 1235–1243.]

[21] Zhang X, Wu X, He S, et al. Precipitation forecast based on CEEMD- LSTM coupled model[J]. Water Science & Technology Water Supply, 2021, 21(22): 1–17.

[22] 许德合, 丁严, 张棋, 等. EEMD-ARIMA在干旱预测中的应用——以新疆维吾尔自治区为例[J]. 中国农村水利水电, 2021 (7): 1–11. [Xu Dehe, Ding Yan, Zhang Qi, et al. Application of the EEMD- ARIMA combined model in drought prediction: A case study in Xinjiang Uygur Autonomous Region[J]. China Rural Water and Hydropower, 2021(7): 1–11.]

[23] Tsakiris G, Vangelis H. Towards a drought watch system based on spatial SPI[J]. Water Resources Management, 2004, 18(1): 1–12.

[24] Moreira E E. SPI drought class prediction using log-linear models applied to wet and dry seasons[J]. Physics and Chemistry of the Earth, 2016, 94: 136–145.

[25] Alquraish M, Abuhasel K A, Alqahtani A S, et al. SPI-based hybrid hidden Markov-GA, ARIMA-GA, and ARIMA-GA-ANN models for meteorological drought forecasting[J]. Sustainability, 2021, 13(22): 12576, doi: 10.3390/su132212576.

[26] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings Mathematical Physical & Engineering Sciences, 1998, 454(1971): 903–995.

[27] Wu Z, Huang N E. Ensemble empirical mode decomposition: A noise-assisted data analysis method[J]. Advances in Adaptive Data Analysis, 2009, 1: 1–41.

[28] Yeh J R, Shieh J S, Huang N E. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method[J]. Advances in Adaptive Data Analysis, 2010, 2(2): 135–156.

[29] 孙堃, 赵萌萌, 沈美娜, 等. 基于CEEMD和模糊熵的随机森林风力发电功率预测[J]. 智慧电力, 2019, 47(10): 36-43. [Sun Kun, Zhao Mengmeng, Shen Meina, et al. Wind power forecasting with random forest based on CEEMD and fuzzy entropy[J]. Smart Power, 2019, 47(10): 36–43.]

[30] Box G E P, Jenkins G M. Time series Analysis: Forecasting and Control[M]. San Francisco: Holden Day, 1976.

[31] Dilling S, Macvicar B J. Cleaning high-frequency velocity profile data with autoregressive moving average (ARMA) models[J]. Flow Measurement & Instrumentation, 2017, 54: 68–81.

[32] 左秀霞. 带高次趋势项的ADF单位根检验[J].数量经济技术经济研究, 2019, 36(1): 152-169. [Zuo Xiuxia. ADF unit rot test with high order trend term[J]. The Journal of Quantitative & Technical Economics, 2019, 36(1): 152–169.]

[33] Liu Q, Zhang G, Ali S, et al. SPI-based drought simulation and prediction using ARMA-GARCH model[J]. Applied Mathematics and Computation, 2019, 355: 96–107.

[34] Adikari K E, Shrestha S, Ratnayake D T, et al. Evaluation of artificial intelligence models for flood and drought forecasting in arid and tropical regions[J]. Environmental Modelling & Software, 2021, 144(4): 105136, doi: 10.1016/j.envsoft.2021.105136.

Applicability of the LSTM and ARIMA model in drought prediction based on CEEMD: A case study of Xinjiang

DING Yan[1], XU Dehe[1], CAO Lianhai[1], GUAN Xiangrong[2]

(1. College of Surveying and Geo-Informatics, North China University of Water Resources and Electric Power, Zhengzhou 450046, Henan, China; 2. E-Government Center of Natural Resources in Henan Province, Zhengzhou 450046, Henan, China)

Abstract: The frequent occurrence of droughts seriously affects normal agricultural production and economic development. Accurate prediction of drought occurrence is of great importance in reducing drought losses. Nevertheless, drought occurrences have not been well predicted. Drought indices can be used to quantitatively evaluate the intensity, duration, and influence range of drought. Thus, on the basis of daily precipitation data from 1960 to 2019 in the Xinjiang Uyghur Autonomous Region, the standardized precipitation index (SPI) at timescales of 1, 3, 6, 9, 12, and 24 months were calculated. Aiming for the nonlinear and nonstationary characteristics of SPI, a new drought prediction method was proposed combining the single model and the complementary ensemble empirical mode decomposition (CEEMD), which can process nonlinear and nonstationary signals. In this paper, the autoregressive integrated moving average (ARIMA) model, the long short-term memory (LSTM) network, the CEEMD-ARIMA combined model, and the CEEMD-LSTM combined model were constructed to predict a multiscale SPI. The validity of prediction models was determined using root mean square error, mean absolute error, and coefficient of determination (R^2). Kriging interpolation was used to demonstrate the predicted results of the four models. The results revealed that the forecast accuracy of the four models increases with the increase of SPI timescales, and the highest accuracy is obtained at SPI24. CEEMD decomposition can effectively stabilize the time series. Drought prediction based on the CEEMD provides a stable premise for the single model. At each timescale, combined models obtain higher prediction accuracy than single models, which indicates that combined models are more suitable for drought prediction. The forecast accuracy of the four models in order from the lowest to highest accuracy is the LSTM model, followed by the ARIMA, CEEMD-LSTM, and CEEMD-ARIMA models (the maximum R^2 values are 0.8882, 0.9103, 0.9403, and 0.9846, respectively). The CEEMD-ARIMA model shows the best ability to forecast SPI values. This study explored the applicability of four drought prediction models and provided a basis for meteorological disaster prevention and mitigation efforts.

Keywords: complementary ensemble empirical mode decomposition; long short-term memory network; autoregressive integrated moving average; standardized precipitation index; drought prediction; Xinjiang